細粒度動的負荷分散機構を備えたネットワーク・スーパーコンピューティング環境の構築

AUTHOR(S):

富田, 眞治

細粒度動的負荷分散機構を備えたネットワーク・スーパーコンピューティング環境の構築

課題番号　12558027

平成12年度～平成13年度科学研究費補助金 基盤研究(B)(2)
研究成果報告書

平成１４年３月

研究代表者 富田 眞治
(京都大学 大学院情報学研究科 教授)

平成 12 年度～平成 13 年度 科学研究費補助金 基盤研究 (B)(2)
研究成果報告書

1. 課題番号 12558027

2. 研究課題 細粒度動的負荷分散機構を備えた
ネットワーク・スーパーコンピューティング環境の構築

3. 研究組織

研究代表者: 富田 眞治 （京都大学大学院 情報学研究科 教授）
研究分担者: 北村 俊明 （京都大学 総合情報メディアセンタ 助教授）
研究分担者: 中島 康彦 （京都大学大学院 経済学研究科 助教授）
研究分担者: 森 眞一郎 （京都大学大学院 情報学研究科 助教授）
研究分担者: 五島 正裕 （京都大学大学院 情報学研究科 助手）
研究分担者: 津邑 公暁 （京都大学大学院 経済学研究科 助手）

研究協力者: 小西 将人, Damien Le Moal, 増田 峰義
鳥崎唯之, 額田 匡則, 生雲 公啓

4. 交付決定額 (配分額) (金額単位:千円)

	直接経費	間接経費	合計
平成 12 年度	5,600	0	5,600
平成 13 年度	7,000	0	7,000
総 計	12,600	0	12,600

5. 研究発表

(1) 学会誌, 国際会議等

1. 秤谷雅史, 齋藤康二, 小西将人, 五島正裕, 森眞一郎, 富田眞治, 「超並列計算機 JUMP-1 における分散共有メモリ管理」, 並列処理シンポジウム *JSPP200*, pp.67–74, 2000 年 5 月.
2. 五島正裕, 齋藤康二, 小西将人, 秤谷雅史, 森眞一郎, 富田眞治, 「超並列計算機 JUMP-1 の分散共有メモリ・システム」, 情報処理学会論文誌:ハイパフォーマンスコンピューティングシステム, Vo.41, No.SIG 8(HPS 2), pp.15–27, 2000 年 11 月.
3. 小西将人, 五島正裕, 森眞一郎, 富田眞治, 「超並列計算機 JUMP-1 における分散共有メモリ管理の実装」, 情報処理学会論文誌, Vol.42, No.4, pp.674–682, 2001 年 4 月.

4. Damien Le Moal, Mineyoshi Masuda, Masahiro Goshima, Shin-ichiro Mori, Yasuhiko Nakashima, Toshiaki Kitamura and Shinji Tomita, "PRIORITY ENHANCED STRIDE SCHEDULING," *Int'l Conf. on High Performance Computing in Asia-Pacific Region*, Sept. 2001.
(On-line Proceedings at http://www.gu.edu.au/ins/its/hpcasia2001/)

(2) 口頭発表

1. 小西将人, 五島正裕, 森眞一郎, 富田眞治, 「超並列計算機 JUMP-1 における分散共有メモリシステムの性能評価」, 情報処理学会 研究会報告, 2000-ARC-139, pp.19–24, 2000 年 8 月.
2. 増田峰義, 鳥崎唯之, 五島正裕, 森眞一郎, 富田眞治, 「分散 OS Colonia における並列アクティビティの高速移送」, 情報処理学会 研究会報告, 2000-OS-85, pp.23–30, 2000 年 8 月.
3. 額田匡則, 鈴木紀章, 天野英晴, 西村克信, 田村友紀, 長名保範, 小西将人, 五島正裕, 富田眞治, 「超並列計算機 JUMP-1 のマルチキャスト機構による性能向上」, 電子情報通信学会技術研究報告, CPSY2001-49, pp.37-44, 2001 年 7 月.
4. 小西将人, 額田匡則, 五島正裕, 森眞一郎, 富田眞治, 「超並列計算機 JUMP-1 の性能評価」, 情報処理学会 研究会報告, 2001-ARC-144, pp.189-194, 2001 年 7 月.

6. 研究概要

我々は, 複数の独立した計算機が必要に応じて群（コロニー）を形成し, それらがあたかも単一の計算機であるかのようなイメージ (Single System Image:SSI) を提供することで従来の SMP システムの操作性に匹敵する計算機環境を実現するコンピュータ・コロニーの研究を行ってきた. 「ネットワークの向こうにある無尽蔵の計算資源を, その物理的な構成を意識せずに自由に使いこなせる計算機環境」がコンピュータ・コロニーの理想である. 我々はこの目的を達成するために必要な共有メモリ環境を効率的に実現するハードウェアと, プログラミングフェースとして従来の SMP システムでのタスク-スレッド・モデルを拡張したミッション-ユニット・モデルの基礎的な研究を行ってきた.

本研究の目的は, 今までの研究成果を統合しネットワーク・スーパーコンピュータの一形態としての**コンピュータ・コロニー**をプロトタイプシステムとして実現することである.

主な研究成果は以下の通りである.

負荷分散機構の実装と評価　　複数計算機間での細粒度動的負荷分散を実現するグローバル・スケジューラの戦略を決定する際に必要なシステムパラメータを同定するため, 負荷分散機構において実質的な作業を行うユニット移送機構, ホームページ移送機構の開発を行った.

ユニット移送機構の開発 共有メモリ環境であることを活用し, ユニット移送時には必要最低限の情報のみを移送することで, 移送を決定後から移送先での実行再開までの時間を短縮する基本プロトコルを設計し, それに基づいてたベーシックな実装を行った. その結果, 移送先でのオンデマンドなメモリ領域確保／割り当ての時間がユニット移送において極めて問題であることが判明した. そこで, ユニット移送機構自身が独自に管理するメモリ領域を一定量だけ予め割り当てる実装方式を提案した. さたに, 一旦移送されたユニットが再び移送前のノードに戻ってくる場合に, 過去の履歴を利用することで不必要なデータの移送を軽減する差分移送方式を提案し, 具体的な実装方法について検討を行った.

ホームページ移送機構の開発 ユニットの移送に伴うホームページの移送ならびに, 実行されるプログラムの振る舞いに応じて, プログラムから明示的／非明示的にホームページを移送するための基本プロトコルを作成しその実装を行った. 頻繁に利用され, かつ, タイミングがクリティカルな一部機能を専用の通信ハードウェアの機能を利用して実装することで, プロトコルプロセッサの負担を大きく軽減できることを確認した.

共有メモリ環境を提供する通信ハードウェア上のプロトコル開発 平成 10 年度からの基盤研究の成果として作成したネットワーク・インタフェース・カード上に搭載された FPGA の詳細設計を行った. 具体的には, 共有メモリの一貫性制御を支援するプロトコルプロセッサインタフェース, ワークステーションに実装するためのバスインタフェース, 1Gpbs の光接続部の基本インタフェースを設計した. この通信ハードウェアを, ワークステーション上のプログラムの仮想アドレスにマッピングし, 2 台のワークステーション間で光接続により通信を行うためのプロトタイプハードウェア環境を構築した. その結果, 光接続部におけるデータの誤り率が予想以上に高いことが判明した. 解析の結果, 単体でのループバック試験に関しては全く問題がおこらず, 複数台構成のシステムにおいてはじめて問題が発生することが分かった. しかし, 根本的な原因の解析には最新の測定器等が必要であり予算ならびに研究期間の問題から, 原因の解析は保留し, この物理層レイアでの問題を上位レイアのプロトコルで回避して, 高いエラー率の下でも通信品質を確保するプロトコルを開発した. また, このプロトコルを FPGA 上のハードウェアで実装した. その結果, 上位レイアでのエラー発生率を 1/5 に軽減し, かつエラーフリーの Read ならびに Write アクセスを, それぞれ 3.7 ならびに 4.1 マイクロ秒 (平均) で実現できることがわかった.

コンピュータ・コロニーの実装と性能評価

プロトタイプ環境として, SPARC 版 LINUX オペレーティングシステムを搭載した 2 台のワークステーションに独自開発の専用ネットワークインタフェースカードを接続した簡易共有メモリ環境を構築し, その性能評価を行った. 専用カードをユーザレベルで利用可能とするためのドライバ開発ならびに, Linux カーネルの修正を行った. ネットワークカードが提供する共有メモリへのアクセス時間

の評価を行った結果, 細粒度のローカルメモリアクセスに関して, ユーザレベルの直接アクセスはシステムコールを介した場合に比べて, 約 8 倍の高速化が得られることがわかった. また, リモートメモリアクセスの平均レイテンシは, Read アクセスで 2.7 マイクロ秒, Write アクセスで 2.3 マイクロ秒となり, プロトタイプシステムとしては満足な性能が得られた.

目次